DOCUMENT RESUME

ED 395 973 TM 025 111

AUTHOR Martinez, Michael E.
TITLE A Comparison of Multiple-Choice and Constructed Figural Response Items.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-90-19
PUB DATE Oct 90
NOTE 39p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Constructed Response; Difficulty Level; Elementary Education; *Elementary School Students; Field Tests; Guessing (Tests); High Schools; *High School Students; *Multiple Choice Tests; Research Needs; Sciences; *Test Items
IDENTIFIERS *Figural Response Items; National Assessment of Educational Progress

ABSTRACT

In contrast to multiple-choice test questions, figural response items call for constructed responses and rely upon figural material, such as illustrations and graphs, as the response medium. Figural response questions in various science domains were created and administered to a sample of 347 fourth, 365 eighth, and 322 twelfth graders. Data were gathered in conjunction with field testing for the 1990 National Assessment of Educational Progress. Item and test statistics from parallel sets of figural response and multiple-choice questions were compared. Figural response items were generally more difficult, especially for questions that were difficult ($p<.5$) in their constructed-response forms. Figural response questions were also slightly more discriminating and reliable than their multiple-choice counterparts, but had higher omit rates. The paper addresses the relevance of guessing to figural response items and the diagnostic value of the item type. Plans for future research on figural response items are discussed. An appendix describes figural response items. (Contains 3 tables, 10 figures, and 20 references.) (Author/SLD)

RESEARCH REPORT

A COMPARISON OF MULTIPLE-CHOICE AND CONSTRUCTED FIGURAL RESPONSE ITEMS

Michael E. Martinez

Educational Testing Service
Princeton, New Jersey
October 1990

A Comparison of Multiple-Choice and Constructed Figural Response Items

Michael E. Martinez

Educational Testing Service

Princeton, NJ

Running head: MULTIPLE-CHOICE VS. FIGURAL RESPONSE

Abstract

In contrast to multiple-choice test questions, figural response items call for constructed responses and rely upon figural material, such as illustrations and graphs, as the response medium. Figural response questions in various science domains were created and administered to a sample of 4th-, 8th-, and 12th-grade students. Item and test statistics from parallel sets of figural response and multiple-choice questions were compared. Figural response items were generally more difficult, especially for questions that were difficult ($p<.5$) in their constructed-response forms. Figural response questions were also slightly more discriminating and reliable than their multiple-choice counterparts, but had higher omit rates. The paper addresses the relevance of guessing to figural response items and the diagnostic value of the item type. Plans for future research on figural response items are discussed.

A Comparison of Multiple-Choice and Constructed Figural Response Items

Testing has been under the public spotlight with increasing regularity (Fiske, 1990). Tests are found to have numerous shortcomings; most existing tests are not diagnostic in nature, and their relevance to instruction is questionable (Nickerson, 1989). But more generally, standardized test questions are perceived as failing to elicit a full range of desirable cognitive processes, which for many educators include thinking critically, synthesizing ideas, and formulating and carrying out plans (Frederiksen, 1984; Haney & Madaus, 1989; Wiggins, 1989).

Multiple-choice items are frequently the target of this criticism, in part because they are commonly thought to require no more than recall of information. Reacting to this dissatisfaction, researchers in testing organizations, state departments of education, universities, and research institutions have tried to develop alternatives to multiple-choice tests. The products of these efforts include performance- and portfolio-based assessments.

This paper describes an additional alternative to multiple-choice: figural response items. Figural response differs from multiple-choice in two ways. First, figural response items call for constructed responses -- answers made up in the mind of the examinee rather than chosen from a list of options (Cronbach, 1984). Constructed-response items are sometimes referred to as free response or open-ended items, and finer distinctions among these terms can be found (c.f. Birenbaum & Tatsuoka, 1987; Ward, Frederiksen, & Carlson, 1980). A second feature of figural response items is their dependence upon

figural material, such as illustrations and graphs, as the response medium. The examinee responds to questions or directions by carrying out operations on a figure. For example, given a diagram of a heart, an examinee might be asked to identify a specific valve or point out an anatomical flaw in the diagram, both of which could be done by marking a location on the figure. Other figural response items ask examinees to indicate direction by using free-form arrows or to show relationships between variables by drawing a line graph. In this paper, figural response items should be understood to be a kind of constructed response item. Examples of figural response items used in this study are shown in Figures 1-9.

Insert Figures 1-9 about here

This study contrasted multiple-choice and constructed-response items. In related research (cf. Traub & MacRury, in press) two parallel forms of a test are constructed, one using multiple-choice questions, the other using constructed-response questions that are stem equivalent. Stem equivalence means that item stems (questions or instructions) are identical, but the response options provided in the multiple-choice format are eliminated in the alternative format.

Contrasts between constructed-response and multiple-choice questions have not yielded unambiguous conclusions, but some patterns have emerged. In general, constructed-response questions are somewhat more difficult and reliable than stem-equivalent multiple-choice counterparts (Traub & MacRury, in press). A possible explanation for this finding is that the

probability of guessing correctly on a multiple-choice question is non-trivial, since the number of response options is limited. Guessing correctly introduces error into test scores, thus lowering reliabilities and associated parameters. These effects are reflected in applications of three-parameter item response theory, in which there is a limited capability to estimate ability in lower ability ranges. This degradation in accuracy might be largely attributable to guessing.

In answering multiple-choice questions, examinees might also capitalize on abilities that have little to do with the construct being measured. For example, researchers have identified a response elimination strategy whereby examinees eliminate implausible distractors, and then guess from the remaining options (Snow, 1980). Unless the target construct is intended to embrace such strategizing, error is introduced because the items discriminate on the basis of abilities that lie beyond the pale of the construct.

Method

Subjects

Subjects from grades 4 (N= 347), 8 (N= 365), and 12 (N= 322) were drawn from a national sample of students representing a broad range of characteristics, such as racial/ethnic group, socioeconomic status, and national region of residence. Data were gathered in conjunction with field testing for the 1990 National Assessment of Educational Progress (NAEP).

Items

Twenty-five constructed-response items were written in three content areas: life sciences, physical sciences, and earth/space sciences. The questions were developed in accordance with NAEP content and process

specifications (National Assessment of Educational Progress, 1988). Each question was classified by topic (e.g., ecology), nature of science (e.g., designing an experiment), and thinking skill (e.g. solving problems). The items were reviewed by an outside panel of scientists.

Each figural item was matched with a multiple-choice counterpart, some of which were already part of the NAEP item pool. Twenty-three of the items had four response options each; the remaining two items offered five response options. Other than stem differences needed to clarify the intended response (e.g., "draw an X" or "draw arrowheads"), wording was parallel across items. Descriptions of all 25 figural response items and the overlap of items across grade levels are provided in the Appendix.

The items on which this paper are based were part of a larger field test of science items. Subjects were given three blocks of science items and were allowed 15 minutes per block. The figural response items composed an entire block while the multiple-choice items, which typically are answered more quickly, were accompanied by ancillary multiple-choice science items not connected with the study. The blocks associated with the present project were placed last in the groups of three. The assignment of subjects to condition (figural response or multiple-choice) was random.

Scoring

Procedures for classifying responses, or scoring rubrics, were developed for constructed-response items. For some items, response categories defined by the rubrics separated different kinds of conceptual errors. For example, on an item that called for prediction of an object's trajectory, the rubrics included categories for common misconceptions about the object's

path. For other items, categories reflected ordinal counts of correct responses. On one question, for example, students were shown a map of the earth's geological plates. To show which way the plates moved, students drew arrowheads on shafts that were provided. The scoring rubric for that item reflected the number of arrowheads placed correctly.

In all but four items, multiple-choice response options had parallel response categories in the constructed-response versions. Scoring rubrics specified a single correct answer for each item, and total scores were based on the number of items answered correctly. When item statistics were computed, the multiple scoring categories based on the rubrics were reduced to correct/incorrect judgments for both figural response and multiple-choice items. Most constructed responses were hand-scored once by one of two graders. A random subset of items was scored twice and showed inter-rater reliability (Cohen's Kappa) values of 0.80, 0.77, and 0.80 for grades 4, 8, and 12, respectively. These reliabilities are based on discrepancies between scorers on all response categories; reliabilities are likely to be higher for correct/incorrect judgments.

Responses to multiple-choice counterparts of constructed response questions were key-entered. In both item formats, statistics were derived for item difficulty, item/total score relationships, reliabilities, standard errors of measurement, and non-response rates.

Results

Item Difficulty

Across grade levels, constructed-response questions were, in general, more difficult than their multiple-choice counterparts (Table 1). Upon closer

Insert Table 1 about here

examination, it became evident that the relative difficulty between formats interacted with the difficulty of the question in its constructed-response form. For questions that were relatively difficult (p less than or equal to .5), constructed response questions were almost uniformly more difficult: 29 out of 33 were more difficult as constructed-response items (sign test, $p<.001$). But for items with p greater than .5, 9 of 11 were more difficult in their multiple-choice format (sign test, $p<.05$). In these calculations, items that were administered at more than one grade level were counted separately at each grade level.

Figure 10 is a scatterplot of item difficulties in two formats. The plot

Insert Figure 10 about here

shows a few potentially important patterns. First, most of the data points fall above the diagonal, illustrating that, in general, multiple-choice items were easier than constructed-response items. Divergence from the diagonal is greater with the more difficult questions. As items become easier, the distribution converges toward the diagonal, and differences in difficulty between formats drop off. The empty lower right-hand corner of the plot shows that even when multiple-choice items are more difficult than constructed-response items, these differences in difficulty are not great. Finally, only a few of items were very difficult ($p<.20$) in their multiple-choice

formats, whereas a number of items were very difficult in their constructed-response formats.

Item/Total Score Relationships

Item/total score (r-biserial) correlations were generally higher for constructed-response items than for their multiple-choice counterparts (Table 2). Each of these correlations used as its criterion score a set of items of the

---

Insert Table 2 about here

---

same format (e.g., multiple-choice items were used to predict a multiple-choice total score). At the three grade levels tested, the constructed response items were better predictors of their own total score than were multiple-choice items.

The superior discrimination offered by constructed-response items was moderated by whether they were easier or more difficult than their multiple-choice counterparts. Where the constructed-response versions were easier, format differences in discrimination were small. Mean discrimination values for these items were as follows: at Grade 4, 0.62 for constructed-response and 0.65 for multiple-choice; at Grade 8, 0.56 for constructed-response and 0.51 for multiple-choice; at Grade 12, 0.53 for both constructed-response and multiple-choice. Where constructed-response items were more difficult, advantages in discrimination for the constructed-response format were more evident. Mean discrimination values were: at Grade 4, 0.62 for constructed-response and 0.41 for multiple-choice; at Grade 8, 0.49 for constructed-response and 0.43 for

multiple-choice; at Grade 12, 0.60 for constructed-response and 0.42 for multiple-choice.

Reliabilities and Standard Errors of Measurement

At grades 8 and 12, the reliabilities for total scores were marginally better in the constructed-response format (Table 3). The standard errors were uniformly and markedly lower for the constructed-response items across grades.

Insert Table 3 about here

Response Patterns

For most constructed-response items, categories corresponding to the multiple-choice options captured most of the responses. But in some cases, the constructed response format (a) resulted in responses that had different diagnostic implications, or (b) elicited substantially different distributions of responses. These points are illustrated by the "heat & temperature" problem, in which a temperature x heat graph is presented (Fig. 1). The temperature axis ranges from -30 to 120 degrees Celsius, and the relationship between heat and temperature is plotted up to 42 degrees. The constructed-response task is to complete the graph, while the multiple-choice task involves selecting the appropriate completed graph.

When the item was given to 12th-grade students, 43% of the constructed responses corresponded to the four multiple-choice options, including 9.7% that matched the correct option. Two additional response categories had no multiple-choice counterparts. One pattern, drawn by 9.7% of respondents,

showed the temperature/heat slope diminishing in steepness as more heat was applied to water. In the second pattern, 2% of respondents showed the temperature flattening at 100 degrees Celsius, while it should have risen beyond 100 degrees once all water had turned to steam. In addition, the distributions of responses varied according to the format. For example, 14.9% of the respondents chose the multiple-choice option indicating a curvilinear relationship between temperature and heat, but only 2.2% actually drew a curvilinear relationship on the constructed-response version. Finally, 19.6% were able to select the correct answer from four options, but only 9.7% were able to draw the correct response. Across all items and grades, 63% of the constructed responses corresponded to multiple-choice options; 39.8% matched the correct option.

Omitted Items

Items were classified as omitted when at least one item that followed was attempted. Across grade categories, there was a higher incidence of omission for constructed response than for multiple-choice items. At grades 4, 8, and 12, respectively, 4.4%, 2.1%, and 3.2% of multiple-choice items were omitted. Omissions were more frequent for constructed-response items: 6.8%, 4.6%, and 5.9%.

Uncategorized Responses

One characteristic of constructed-response questions is that the full range of answers cannot be pre-specified. It is always possible that an examinee will respond to a question in some unique, rare, or idiosyncratic way. In this study, the number of uncategorized responses was non-trivial: 29.1% at grade 4, 21.4% at grade 8, and 18.1% at grade 12.

The uncategorized responses seemed to be of three types:

1. Random responses, in which no information about the examinee's knowledge could be extracted.

2. Reasoned responses for which there were no defined categories.

3. Responses indicating misinterpretation of the question.

Apparent misinterpretation of the question was evidenced by responses to some of the constructed response questions. In the clearest example of misinterpretation, a ball is shown rolling through a C-shaped tube. On the figure, next to the tube, is the label "Tabletop." The desired response is that, once the ball emerges from the tube, it will travel in a straight line. However, several examinees drew paths in which the ball traveled straight, reversed, and returned toward the tube. This would have been the correct response if the tube were pointed straight up. Responses of this sort were categorized as indicating misinterpretation of the question.

Discussion

This study is a first step in investigating the properties of figural response items, a type of constructed-response test item that, to this point, had not been contrasted with multiple-choice. The item and test statistics reported here show the constructed-response items to be generally comparable or superior to parallel multiple-choice items. The findings corroborate existing research on differences between constructed-response and multiple-choice items (Traub & MacRury, in press), and they raise additional research questions.

In general, constructed-response items were more difficult than multiple-choice items, but differences in difficulty interacted with the

relative difficulty of the items. The common finding that constructed response items are more difficult than multiple-choice needs to be qualified -- it holds true for more difficult, but not for easier items. Other comparisons of difficulty across formats might examine whether a similar pattern emerges.

Differences in difficulty across formats cannot be accounted for by guessing alone, since multiple-choice items are not always easier than constructed-response counterparts. Moreover, when multiple-choice items are easier, their p-values are often not as high as would be expected if all subjects guessed when they did not know the answer. Furthermore, format differences in difficulty cannot be explained by the practice of counting nearly correct responses as incorrect. This kind of judgment applies to only two items, in which ordinal counts of arrowheads placed correctly on shafts determined the category of the response. All arrowheads had to be placed correctly in order for the response to be marked as correct. But even when nearly correct responses (6 out of 7 for one item; at least 10 out of 12 for the other) are counted as correct, the constructed response version of the item is more difficult than its multiple-choice form.

Because constructed-response items spanned more of the difficulty range, they may be well suited to discriminating among high-ability examinees. This characteristic has practical importance, given the difficulty of writing good items that are very difficult. The sparsity of very difficult items ($p<.20$) among multiple-choice items can probably be accounted for in large measure by a guessing factor associated with multiple-choice. Likewise, correct guessing might lead to systematic error, resulting in the reduced reliabilities and discrimination values found for multiple-choice in this and

other studies (Traub & MacRury, in press). One should not overlook alternative explanations for measurement error associated with multiple-choice items, such as their drawing upon strategies that are not directly connected to the construct being measured.

Constructed-response items were found to be generally more discriminating than multiple-choice counterparts. One plausible explanation is that the constructed-response format eliminates random guessing and the error associated with it. However, as previously noted, constructed-response items are sometimes easier than parallel multiple-choice items, implying that the differences cannot be explained solely by a guessing factor. For most items, the constructed-response form is more difficult; this difference might be ascribed, in part, to the ability to guess correctly if that answer is unknown. Among these items, the predictive advantages of constructed-response items are more pronounced, implying that, where guessing is a potential factor, the multiple-choice version loses some ability to discriminate.

Superior item statistics might be a selling point for constructed response items. However, better prediction at the item level might have to be weighed against slightly greater response times for constructed-response questions. In this study, figural response items had statistical advantages even their scoring was less reliable than scoring of multiple-choice counterparts. Automatically scored multiple-choice questions have virtually perfect scoring reliability, but scoring is likely to remain less than perfect for most constructed response questions. Improved reliability of scoring for constructed response formats might heighten their statistical advantages over multiple-choice, not to mention potential advantages in enhanced validity.

Constructed responses were found to have both different diagnostic implications and different response distributions than multiple-choice options. In the "heat & temperature" problem, the constructed-response categories that had no multiple-choice counterparts had different diagnostic meaning and would lead to different instructional prescriptions than did those categories corresponding to multiple-choice options. Furthermore, some multiple-choice options were much less likely to be constructed than chosen. The "heat & temperature" problem thus illustrates that the inferences we draw about a student's understanding might vary according to the assessment format we use.

Omit rates were higher for constructed-response questions. The novelty of the format is a possible explanation: perhaps the unfamiliarity of the test question format caused some examinees to hesitate in providing answers. It also seems reasonable that uncertain examinees would be more likely to attempt an answer to a multiple-choice question, since "the answer is there somewhere," but less likely to construct a tentative response, since the response universe for a constructed-response question is often large as well as unstated. The implications of non-response might be positive. If, in general, students who do not know the correct answer do not respond, correct responses due to guessing are eliminated, possibly resulting in a more accurate estimate of proficiency. On the other hand, uncertainty might make some students reluctant to hazard a response, even if they are correct in their thinking.

The phenomenon of question misinterpretation was clear from constructed responses. For example, in the ball-and-tube question described

earlier, it was clear that some examinees believed that the ball was emerging from a vertical tube pointed upward rather than from a horizontal tube onto a table top. Misinterpretations, though inevitable with multiple-choice questions, are not evident on the basis of the responses, since foils are not intended to detect misinterpretations. But there may be a more important reason for misinterpretation of constructed response questions. Although a multiple-choice question stem might be ambiguous, the response options might clarify the intent of the questions. In this case, the phrase used to pose the question and the choice of responses together show what is asked and the type of response expected. No such help is possible with constructed-response questions. In a later version of the ball-and-tube question, the table and tube are shown from an oblique angle in a corner of the page, clearly indicating the flat orientation of the tube. This sort of explicitness may be needed as a rule for constructed-response questions.

Conclusion

This study focused on item format comparisons at the level of basic item and test statistics. Many other contrasts are possible and are needed, including the extent to which different item formats draw upon different abilities. These studies may be factor-analytic in nature (e.g. Ward, 1982) or may follow an aptitude-treatment interaction (ATI) paradigm (Snow & Lohman, 1989). One could examine whether different formats interact with traits such as test anxiety (Crocker & Schmitt, 1987), whether strategies for various formats differ in the cognitive components needed for solutions (Sternberg, 1982), or whether formats differ in their diagnostic value (Birenbaum & Tatsuoka, 1987). Traditionally, item format contrasts have depended heavily on factor

analysis. Given that the trend to develop new forms of assessment is likely to continue, there is a need for use of alternative research paradigms, since many of these relatively new techniques can provide perspectives not possible using factor analysis and the more basic descriptive comparisons used in this study.

Possible statistical advantages of constructed-response questions have already been noted. Other advantages might attend the use of one format or another, such as influences on the way information is encoded during study (Traub & MacRury, in press), differential dependence on aptitudes or traits, either in capitalizing on strengths or circumventing weaknesses (Crocker & Schmitt, 1987; Snow & Lohman, 1989), or differences in diagnostic value (Birenbaum & Tatsuoka, 1987). A potential disadvantage is the probable lower scoring reliability associated with constructed-response question, whether scored by humans (as in this study) or by machine (Braun, Bennett, Frye, & Soloway, in press). Disadvantages might be outweighed by advantages, particularly if what is gained is a test that is more representative of the knowledge and skill desired in a domain. This raises the large issue of validity, which is clearly central to meaningful format comparisons, but unaddressed in this study and much extant research.

Other research questions, barely addressed elsewhere, are clearly important to the topic of this study. Carroll (1976, p. 34) pointed out the need to have a better understanding of the instructions for carrying out cognitive tasks and, in particular, "the interaction of the instructions with the task performance." In the case of figural response, the instructions are provided in words, while the task is carried out using a picture. The interactions

between verbal (sentential) representations and figural (diagrammatic) representations might draw upon specific skills or productions (Blystone, 1989) or aptitudes that deal with one's ability to form associations across types of representations (verbal and figural) or to transform information from one representational format to another (Sigel, in press). This interaction between representational formats, and especially between verbal and figural, seems ubiquitous in real-world problem solving.

Along with other innovative testing methodologies, the figural response methodology is evolving. Figural response items are now being developed for computer delivery, which expands the kinds of responses possible, including the assembly of structures and the introduction of time sequence processes, such as cell division, as stimuli. The current project will make format comparisons similar to the ones made here, but will also begin to use other research methodologies, including the investigation of aptitude-format interactions. The project is also exploring the range of responses not possible with paper and pencil formats, as well as the possibility of on-line scoring. If automatic scoring is successful, and the item type is found to offer benefits that are difficult or impossible with multiple-choice, the use of figural response and other constructed-response items might be feasible in large-scale testing. If that happens, then perhaps figural response questions along with other non-standard forms of assessment will feed back to instruction with beneficial effects (Frederiksen & Collins, 1989).

References

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats--It does make a difference for diagnostic purposes. Applied Psychological Measurement, 11, 385-395.

Blystone, R. V. (1989). Biology learning based on illustrations. In W. G. Rosen (Ed.), High school biology today and tomorrow. Washington, DC: National Academy Press.

Braun, H. I., Bennett, R. E., Frye, D., & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.). The nature of intelligence. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Schmitt, A. (1987). Improving multiple-choice test performance for examinees with different levels of test anxiety. The Journal of Experimental Education, 55, 201-205.

Cronbach, L. J. (1984). Essentials of psychological testing. New York: Harper & Row.

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. The New York Times, pp. A1, B6.

Frederiksen, N. (1984). The real test bias. American Psychologist, 39, 193-202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18 (9), 27-32

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. Phi Delta Kappan, 70, 683-687.

National Assessment of Educational Progress (1988). Science objectives: 1990 assessment. Princeton, NJ: Educational Testing Service.

Nickerson, R. S. (Ed.). (1989). New directions in educational assessment [Special issue]. Educational Researcher, 18 (9).

Sigel, I. (in press). Representational competence: Another type? In M. Chandler & M. Chapman (Eds.), Criteria for competence: Controversies in the assessment of children's abilities. Hillsdale, NJ: Lawrence Erlbaum Associates.

Snow, R. E. (1980). Aptitude processes. In R. E. Snow, P.-A. Federico, & W. E. Montague (Eds.), Aptitude, learning, and instruction, Volume 1: Cognitive process analyses of aptitude. Hillsdale, NJ: Lawrence Erlbaum Associates.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement (3rd edition). New York: Macmillan.

Sternberg, R. J. (1982). A componential approach to intellectual development. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence, Vol. 1. Hillsdale, NJ: Lawrence Erlbaum Associates.

Traub, R. E., & MacRury, K. (in press). Multiple-choice vs. free-response in the testing of scholastic achievement. To appear in K. Ingenkamp (Ed.), Yearbook on educational measurement. Weinheim: Beltz Publishing Company.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. Applied Psychological Measurement, 6, 1-11.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70, 703-713.

Table 1
Mean Scores (Number Correct) by Item Format and Age Group

Grade Level	Number of Items	Constructed Response			Multiple-Choice			t	d.f.
		n	M	SD	n	M	SD		
4	10	174	2.16	1.36	173	3.82	1.75	9.89*	345
8	15	183	5.25	2.53	182	6.85	2.48	6.10*	363
12	19	160	6.24	3.17	162	8.85	3.40	7.11*	320

*p<.001

Table 2
Mean Discriminations (R-Biserials Correlations)
by Item Format and Age Group

Grade Level	Number of Items	Constructed Response		Multiple-Choice	
		n	Discrimination	n	Discrimination
4	10	174	0.62	173	0.46
8	15	183	0.51	182	0.46
12	19	160	0.57	162	0.45

Table 3
Alpha Reliabilities and Standard Errors of Measurement (SEM) by Item Format and Age Group

Grade Level	Number of Items	Constructed Response			Multiple-Choice		
		n	Reliability	SEM	n	Reliability	SEM
4	10	174	0.35	1.09	173	0.32	1.44
8	15	183	0.63	1.53	182	0.49	1.76
12	19	160	0.71	1.71	162	0.66	1.96

Appendix
Descriptions of Figural Response Items

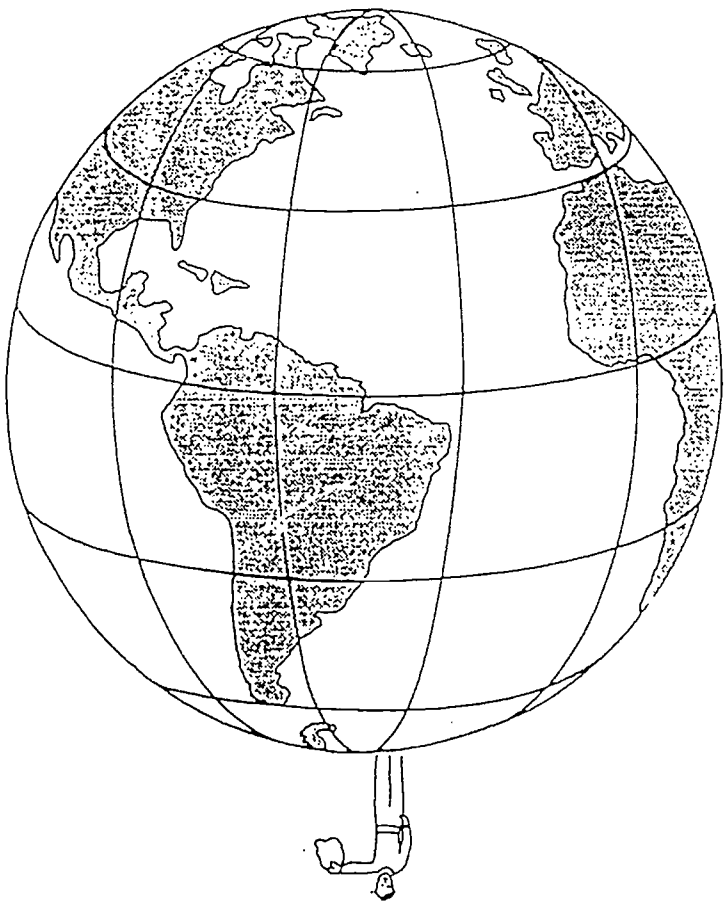
Item Name (Grades)	Item Description (Stimulus and Task)
Globe (4)	Stimulus: Man dropping object at the South Pole. Task: Draw arrow to show direction object will fall.
U-tube Liquid (4)	Stimulus: U-shaped tube with water being poured in. Task: Draw where the water level will settle.
Blood Flow (12)	Stimulus: Diagram of heart and arrow shafts. Task: Indicate blood flow direction with arrowheads.
Evolution (12)	Stimulus: Box labelled "common ancestor" at bottom; At top, labels: turtle, dog, chimpanzee, and human. Task: Draw lines to show evolutionary divergence.
Half-Life (12)	Stimulus: Graph of isotope mass over time. Task: Given half-life, graph radioactive decay.
Heat & temp. (12)	Stimulus: Graph of Temp. x Heat up to 40° C. Task: Complete line graph for 1g. water.
S. Hemisphere (12)	Stimulus: Temperature x month line graph axes. Task: Show seasonal change for S. hemisphere.
Two Species (12)	Stimulus: Graph of species populations over time. Task: Show prey population when predator becomes extinct.
Average Temp. (4, 8)	Stimulus: Graph of temperature by month. Task: Graph the provided temperatures and estimate one temperature.
Spider Legs (4, 8)	Stimulus: Bar graph w/number of legs for bee, horse. Task: Draw bar to show number of legs of a spider.
Eclipse (8, 12)	Stimulus: Diagram of earth and sun, top view. Task: Draw moon in position of solar eclipse.
High Pressure (8, 12)	Stimulus: U. S. map showing barometric pressures. Task: Draw direction of wind flow.

-Continued on the next page-

Appendix, continued

Item Name (Grades)	Item Description (Stimulus and Task)
Mitochondria (8, 12)	Stimulus: Diagram of cell structure. Task: Mark location where cell's energy is produced.
Nucleus (8, 12)	Stimulus: Diagram of cell structure. Task: Mark location where most of DNA can be found.
Plate Tectonics (8, 12)	Stimulus: Earth's geological plates and arrow shafts. Task: Draw arrowheads to show direction of plates.
Retina (8, 12)	Stimulus: Eye viewing object. Task: Draw image of object on retina.
Ball & String (4, 8, 12)	Stimulus: Top view of a weight swung in a circle. Task: Draw trajectory of weight if string is cut.
Food Web (4, 8, 12)	Stimulus: Sun, rabbit, grassland, mouse, and wolf. Task: Draw arrows to show energy flow.
Flatworm (4, 8, 12)	Stimulus: Top view and cross-section of flatworm. Task: On top view, draw a line to show where cross-section was taken.
Glasses Falling (4, 8, 12)	Stimulus: Picture of girl wearing glasses running. Task: Draw path of glasses if girl stops suddenly.
Half-Moon (4, 8, 12)	Stimulus: Diagram of earth and sun, top view. Task: Draw moon in half-moon positions.
Reflected Ray (4, 8, 12)	Stimulus: Light ray striking a flat surface. Task: Draw reflected ray.
Rock Vector (4, 8, 12)	Stimulus: Top view of two persons pulling rock in different directions. Task: Draw vector to show where rock will go.
Thermometer (4, 8, 12)	Stimulus: Celsius thermometer. Task: Fill in mercury column to specified level.
U-tube Ball (4, 8, 12)	Stimulus: Top view of ball rolling in spiral tube. Task: Draw trajectory of ball when it leaves tube.

Below is a diagram of the Earth with a man standing at the South Pole. The man lets go of a rock. Draw an arrow (⟶) that shows in which direction the rock will fall.



NOTE: Person, rock and Earth not drawn to scale.

Figure 1. Globe/Grade 4

On the diagram below, draw where you think the water level would be after all the water in the beaker is poured into the U-shaped tube.

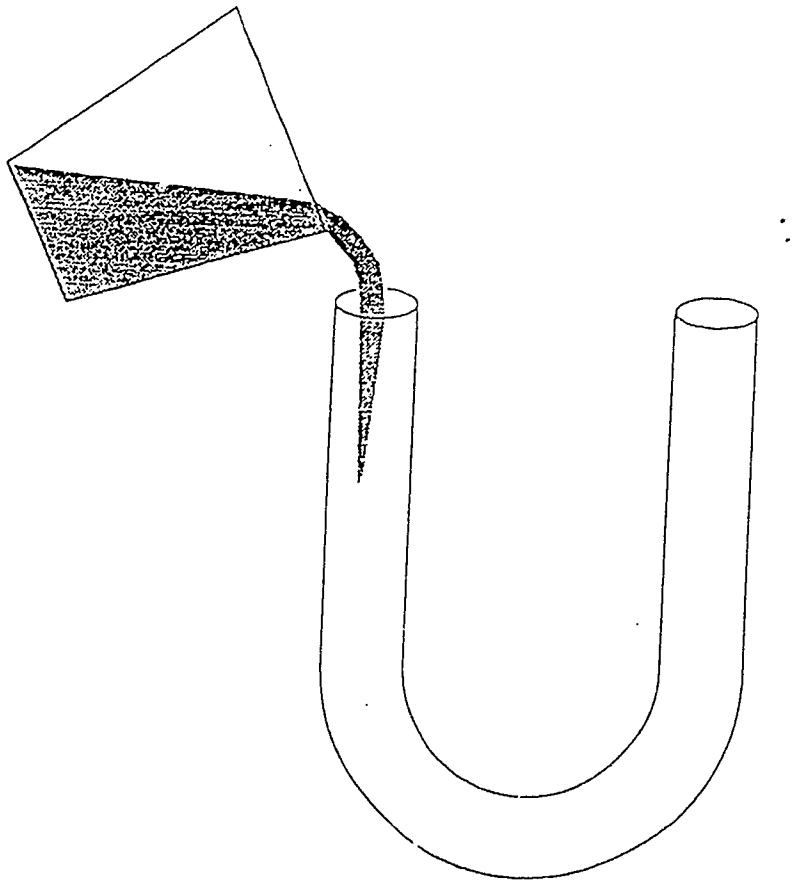
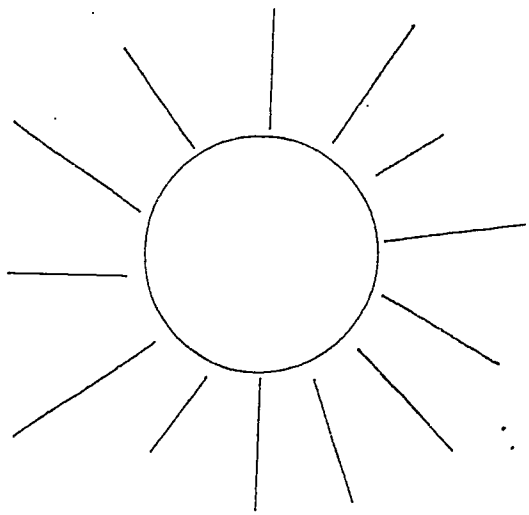
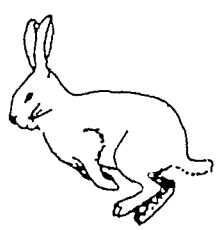
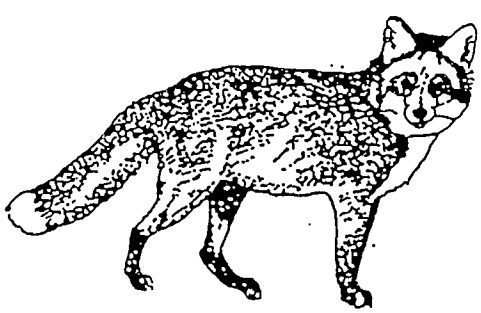


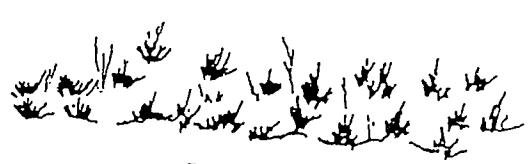
Figure 2. U-tube Liquid/Grade 4

Draw straight arrows to show the directions in which energy flows through the food web shown below.









Grassland

Figure 3. Food Web/Grades 4, 8, 12

In the diagram of the cell shown below, mark an X on the part of the cell where most of the cell's DNA can be found.

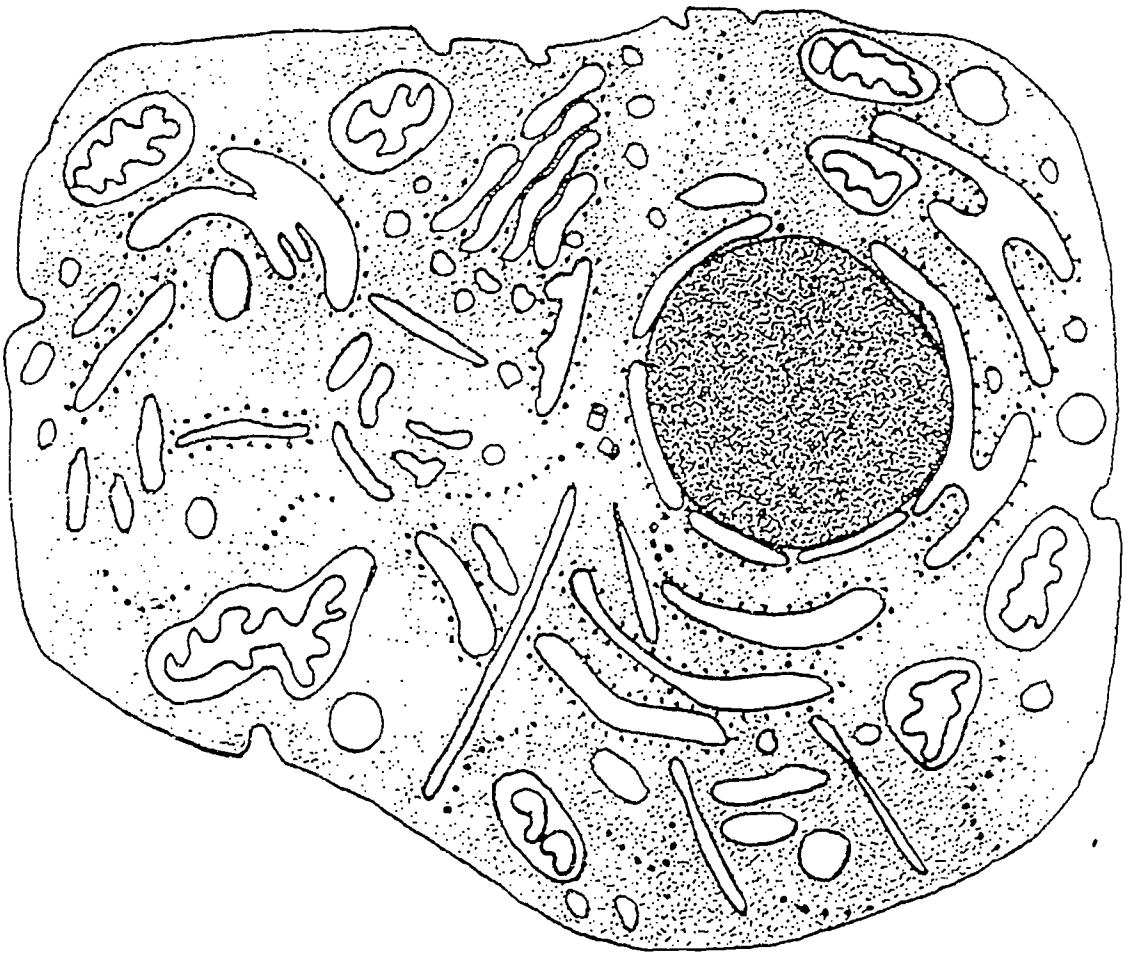


Figure 4. Nucleus/Grades 8, 12

In the diagram of the cell shown below, mark an X on the part of the cell that produces most of the cell's energy as ATP.

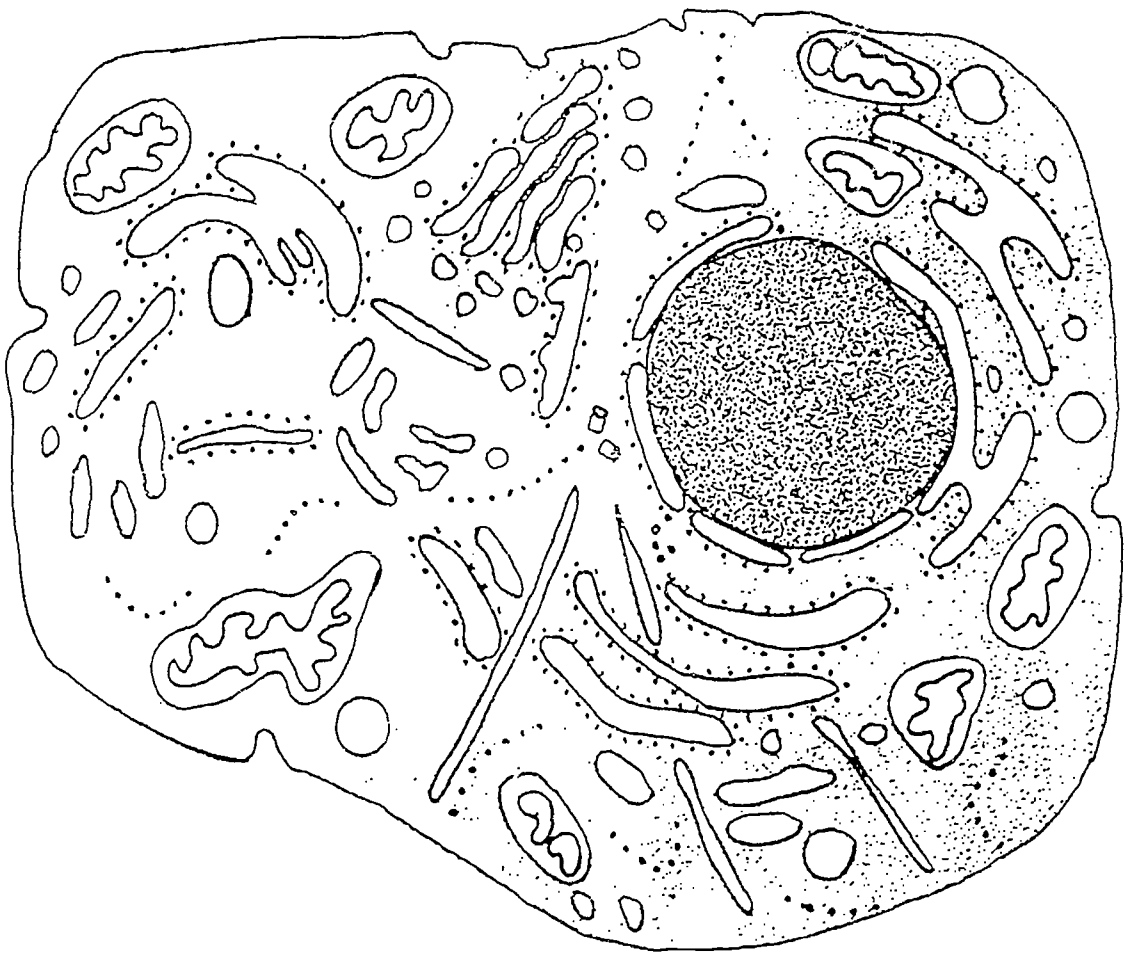


Figure 5. Mitochondria/Grades 8, 12

The map below shows a high-pressure area centered over North Dakota and a low-pressure area centered over Massachusetts. Draw an arrow (⟶) over Lake Michigan that shows the direction in which the winds will blow.

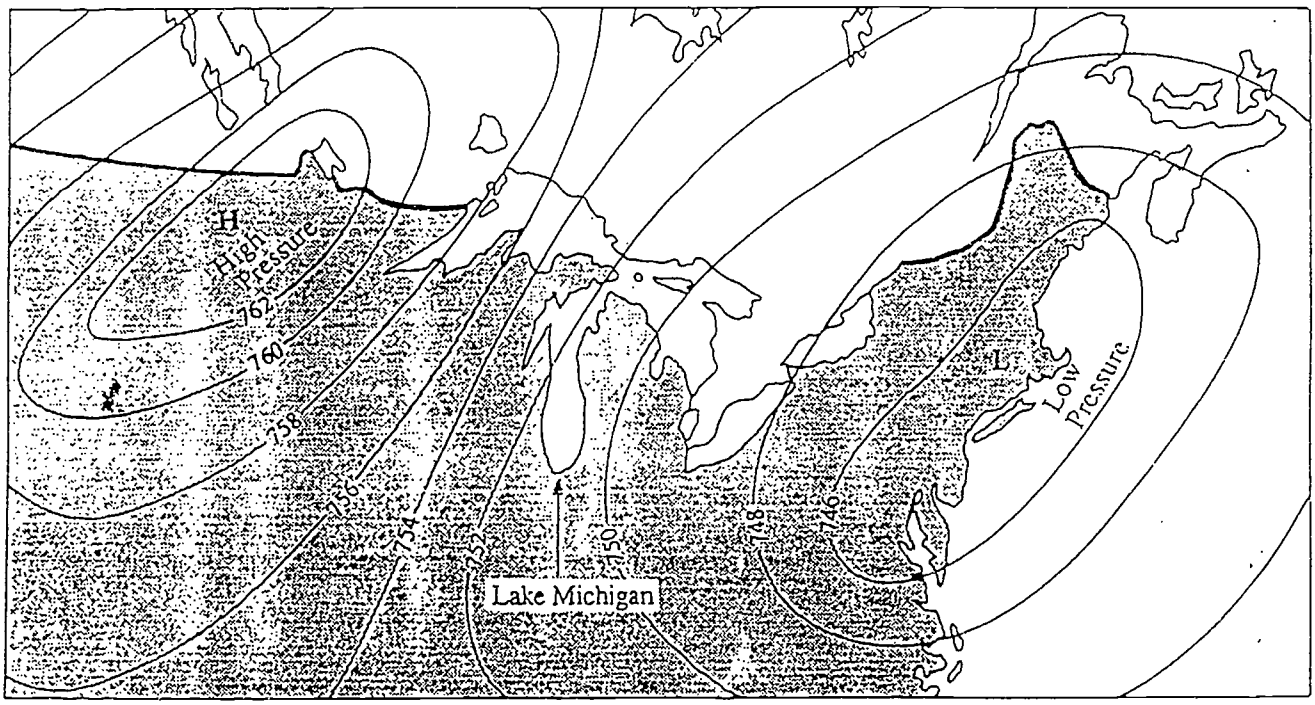


Figure 6. High Pressure/Grades 8, 12

A student started the graph below to show the number of legs of two animals. Complete the graph to show the number of legs that a spider has.

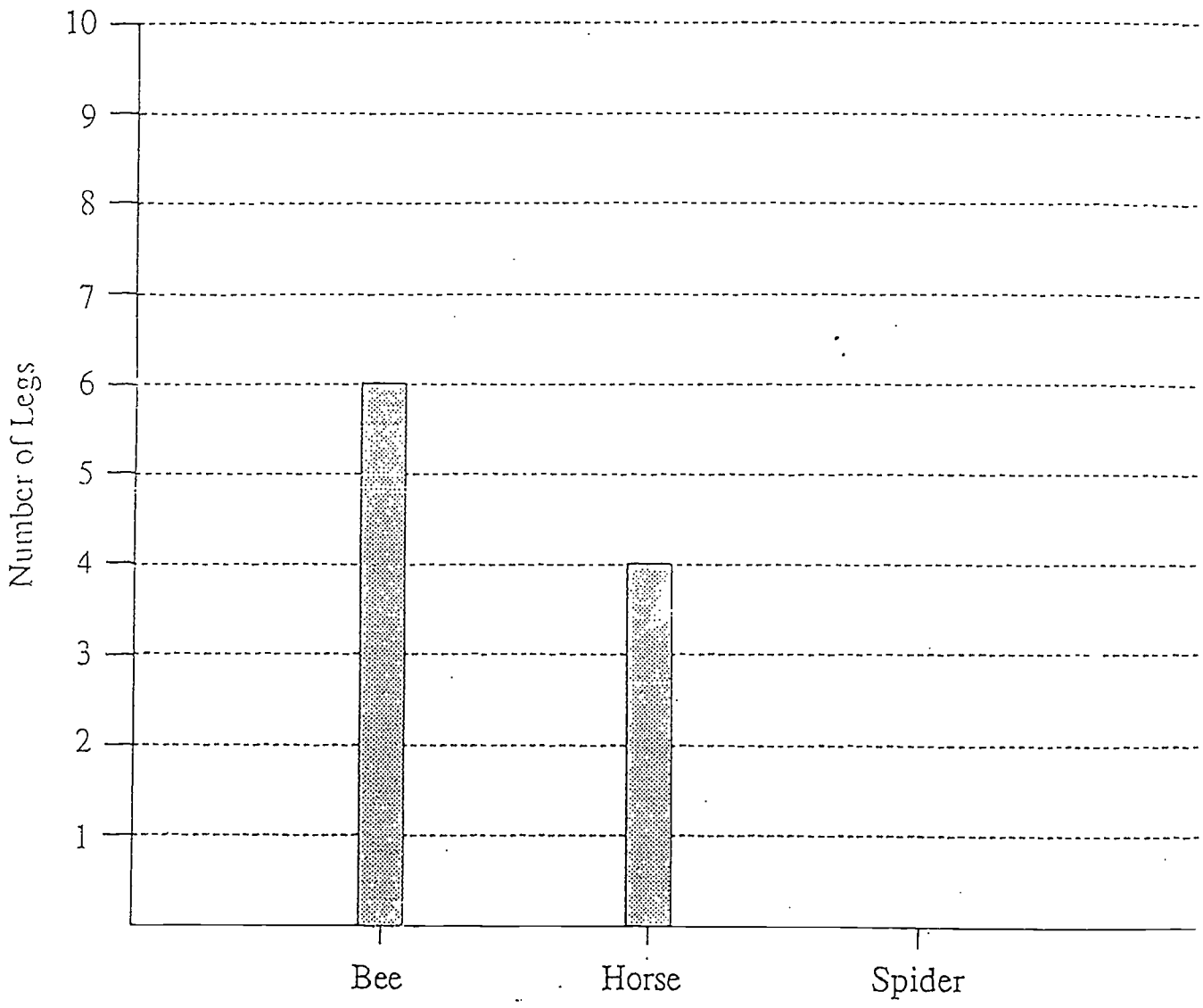
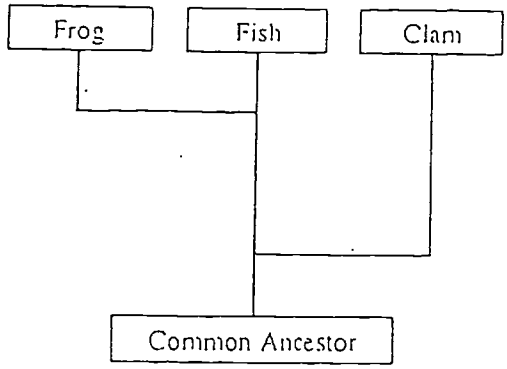


Figure 7. Spider Legs/Grades 4, 8



Draw an evolutionary tree, like the example shown above, that shows the relationship between the following organisms and their common ancestor. Use only straight lines and right ankles.

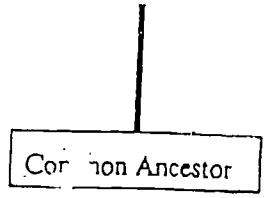


Figure 8. Evolution Tree/Grade 12

The following data show the relationship between the amount of heat added to 1 gram of water and the temperature of that water. When heat is first added, water is in its solid form: ice. Complete the graph that shows how the temperature of water changes with the applied heat.

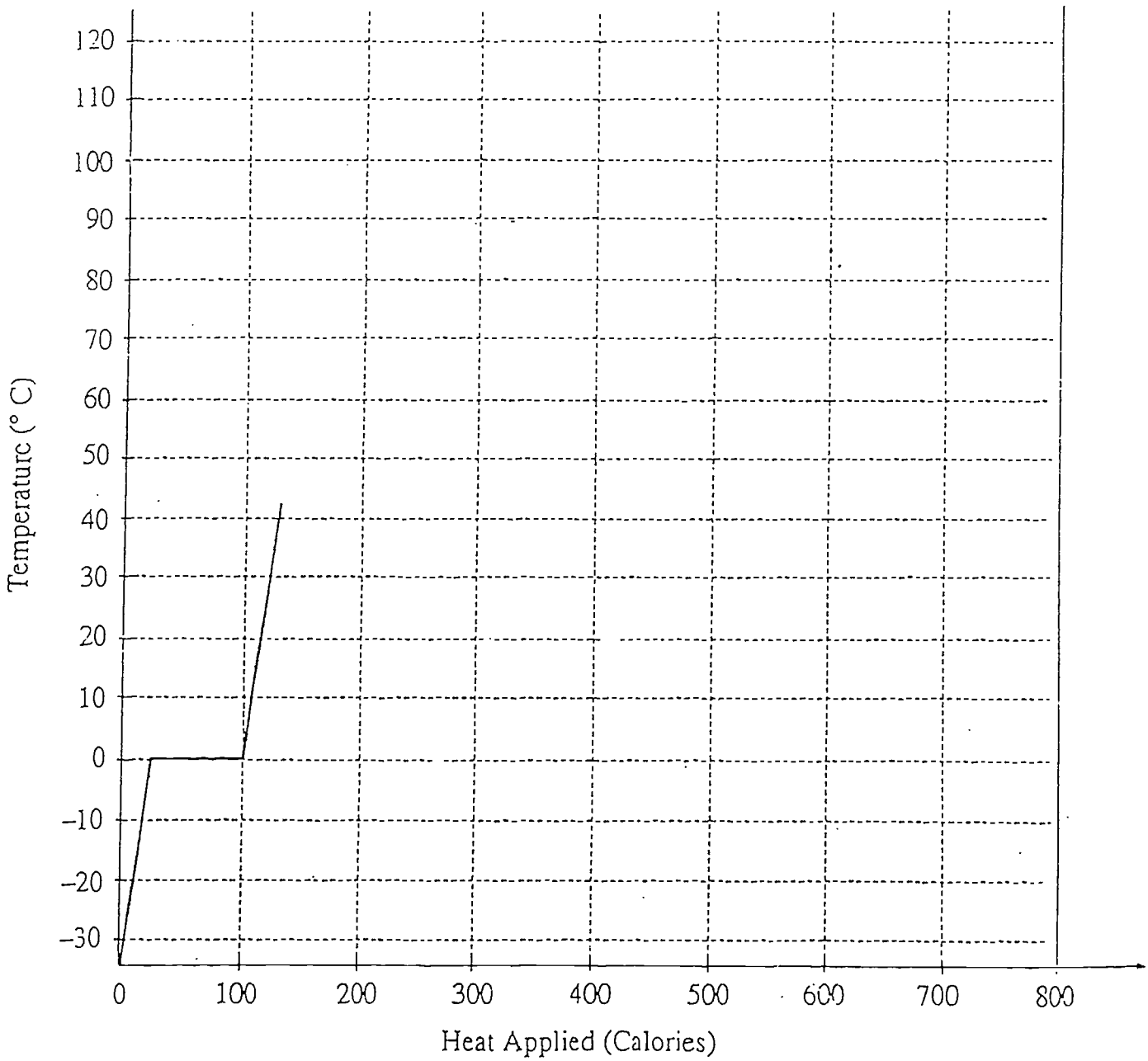


Figure 9. Hot Water/Grade 12

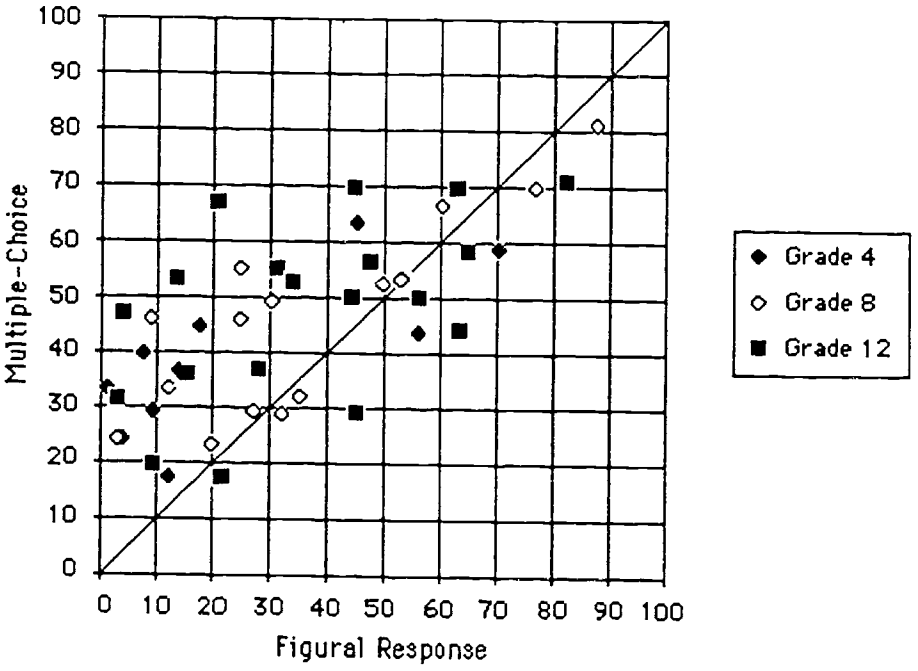


Figure 10. Scatterplot of item difficulties by item format.